Before the

National Telecommunications and Information Administration

Washington, DC

In the Matter of	)	
	)	
"AI Accountability Policy	)	Docket No. 230407-0093
	)	
Request for Comment"	)	

Comment by Kathy Yang, Student, Princeton University

Submitted May 5, 2023

My name is Kathy Yang. I am a student in the Computer Science Department of Princeton University who is also pursuing a certificate through the Center for Information Technology Policy (CITP). I am responding to the NTIA's request for comments on the matter of AI accountability policy. Specifically, I am writing to address Question 22 on "how … the accountability process [should] address data quality and data voids of different kinds."[1] My comment will address these three main points and provide their logical policy takeaways:

1. Data-related concerns surrounding quality and completeness presently exist.
2. There is often a tradeoff between having more complete, transparent data and other important priorities such as privacy and security.
3. Synthetic datasets offer a promising solution for some data quality and completeness problems, but their use in AI must account for their limitations.

Data-related concerns surrounding quality and completeness presently exist.

For years, the notion of "garbage in—garbage out" has dominated the discourse on data quality, referring to the idea that feeding poor data into a computing system will lead to poor outputs. This mantra has persisted in several fields where AI is now being deployed, including healthcare[2] and public sector[3] systems. However, it is important to note that "good" data is just one aspect of many that AI developers should be held accountable for. Furthermore, "good" data is not limited to the quality of the rows that do exist in a dataset—often, AI-related harms occur due to the sparsity or omission of data for a certain subgroup within the dataset.[4] As AI systems

[1] National Telecommunications and Information Administration, U.S. Department of Commerce, "AI Accountability Policy Request for Comment," Federal Register, April 13, 2023, https://www.federalregister.gov/documents/2023/04/13/2023-07776/ai-accountability-policy-request-for-comment#footnote-8-p22433.
[2] Monique F Kilkenny and Kerin M Robinson, "Data Quality: 'Garbage in – Garbage out,'" *Health Information Management Journal* 47, no. 3 (February 2018): pp. 103-105, https://doi.org/10.1177/1833358318774357.
[3] Jan C. Weyerer and Paul F. Langer, "Garbage in, Garbage Out," *Proceedings of the 20th Annual International Conference on Digital Government Research*, 2019, pp. 509-511, https://doi.org/10.1145/3325112.3328220.
[4] Larry Hardesty, "Study Finds Gender and Skin-Type Bias in Commercial Artificial-Intelligence Systems," MIT News (Massachusetts Institute of Technology, February 11, 2018), https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212.

have become more widely adopted, these data-related concerns have likewise proliferated due to factors like the societal biases that pervade real-world data, the relative novelty of large-scale data collection, and the opacity of the current AI industry.

There is a tradeoff between more complete data and other priorities like privacy and security.

From an accountability perspective, we might be tempted to view data transparency and completeness as paramount, since they allow third-party interrogation of data practices. However, there may be legitimate reasons AI developers elect to keep less data and keep data to themselves. The practice of data minimization has gained momentum because it is more cost-effective than data hoarding and reduces the damage associated with security breaches.[5] In regards to privacy, safeguarding sensitive data is an important component of consumer trust.[6] Recent deployment of differential privacy, notably by the U.S. Census Bureau,[7] has lessened the polarity of the transparency-privacy tradeoff. Still, the noise introduced by differential privacy—which obscures information about any particular individual in a given dataset—does reduce the utility of the dataset.[8] From an industry perspective, private actors have the added commercial incentive to protect their data as proprietary in order to obtain market advantages.

Synthetic datasets offer a promising solution for some data problems but possess limitations.

Synthetic data refers to "data that has been generated using a purpose-built mathematical model or algorithm, with the aim of solving a (set of) data science task(s)."[9] The

[5] Bernard Marr, "Why Data Minimization Is an Important Concept in the Age of Big Data," Forbes (Forbes Magazine, March 16, 2016), https://www.forbes.com/sites/bernardmarr/2016/03/16/why-data-minimization-is-an-important-concept-in-the-age-of-big-data/?sh=17097ab51da4.
[6] Brooke Auxier et al., "Americans and Privacy: Concerned, Confused and Feeling Lack of Control over Their Personal Information," Pew Research Center, August 17, 2020, https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/.
[7] John M. Abowd, "The U.S. Census Bureau Adopts Differential Privacy," *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, August 19, 2018, https://doi.org/10.1145/3219819.3226070.
[8] Quan Geng et al., "Privacy and Utility Tradeoff in Approximate Differential Privacy," February 5, 2019, https://arxiv.org/pdf/1810.00877.pdf.
[9] James Jordan et al., "Synthetic Data - What, Why and How?," Royal Society, May 6, 2022, https://royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/Synthetic_Data_Survey-24.pdf.

roots of generating synthetic data can be traced back to the development of Monte Carlo simulation methods in the 1940s, but modern-day approaches often involve deep learning models such as generative adversarial networks (GANs) or variational auto-encoders (VAEs) that are trained on real data.[10] The hope is that generating synthetic data will fill in data gaps and remedy the privacy and fairness concerns associated with real data, where appropriate.

Promising applications of synthetic data

Synthetic data is currently being explored as a tool to enable private data releases, "de-bias" data to improve fairness, and reduce the computational or labor cost associated with labeling data.[11] The first goal of enabling private data releases is especially applicable to the AI accountability process as it would allow researchers and auditors a greater ability to conduct meaningful analyses while still protecting the sensitive real data of a system. The second goal of using synthetic data to remove biases from datasets could be factored into an accountability framework—for example, an auditor might consider whether this practice was used to mitigate racial bias present in an original real-world dataset. Lastly, synthetic data could be used to avoid labor or computation-intensive labeling tasks. An example of this can be seen in computer vision, where the highly impractical pixel-level labeling of shape boundaries can be replaced by generating images with perfect boundaries labeled.[12]

Limitations of synthetic data

Most of the synthetic data being developed for use in AI systems are the product of sophisticated machine learning models themselves. This introduces a "turtles all the way down" scenario, where any diligent accountability policy must consider the quality and completeness

[10] James Jordan et al., "Synthetic Data - What, Why and How?"
[11] James Jordan et al., "Synthetic Data - What, Why and How?"
[12] Sergey I. Nikolenko, "Synthetic Data for Deep Learning," September 26, 2019, https://arxiv.org/abs/1909.11512.

of the training data entered into the models that generate the synthetic data. Because synthetic data must derive its richness from its real data inputs, careful consideration of these inputs is important to avoid obscuring the impurities of original data behind a synthetic mask.

Even though enhanced privacy is a promising application of synthetic data, history encourages caution. There are many examples from the privacy literature where "de-identified" or "anonymized" datasets were released to the public for research and resulted in the compromise of sensitive information.[13] Measures must be taken to ensure that the release of synthetic datasets for the purpose of transparent AI auditing does not come with a similar cost. A particular point of concern is that the privacy properties of synthetic data often fail for outliers, given the tendency for ML models to memorize input data and the sparse nature of outliers.[14] This could lead to the extraction of someone's real information from the synthetic dataset. Differential privacy can be used to mitigate many privacy risks, but its application to synthetic data could yield understudied complications.

Given the above considerations, some policy recommendations are:

1. Hold AI developers accountable for not only the quality of their data, but also for the completeness and transparency of their data.
2. Recognize the tradeoffs between more transparent data and other factors. Simply maximizing transparency is not a good metric and may lead to dangerous breaches of privacy and security.
3. Require developers to reveal some information on the data their AI models are trained on. This disclosure might entail the actual datasets themselves, synthetic proxies of the

[13] Jane Henriksen-Bulmer and Sheridan Jeary, "Re-Identification Attacks—a Systematic Literature Review," *International Journal of Information Management* 36, no. 6 (December 2016): pp. 1184-1192, https://doi.org/10.1016/j.ijinfomgt.2016.08.002.
[14] James Jordan et al., "Synthetic Data - What, Why and How?"

datasets that share important traits with their real counterparts, or higher-level supply chain lists describing the data sources (including clear identification of synthetic ones).

    a. Government standards can help dictate the transparency requirements in various industries depending on the nature of the data common to that industry.
    b. In the interest of privacy, government standards could include a differential privacy requirement for all real and synthetic datasets released.

4. Implement a third-party auditing mechanism to verify that the data released to the public and researchers is actually comparable to the underlying data used in the model.
5. Request that NIST develop and release detailed standards regarding synthetic data generation, release, and auditing based on current best practices.